## Elastic Load Balance

# Guía del usuario

| | |
|---|---|
| **Edición** | 01 |
| **Fecha** | 2025-02-10 |

**HUAWEI CLOUD COMPUTING TECHNOLOGIES CO., LTD.**

## Marcas registradas y permisos

El logotipo HUAWEI y otras marcas registradas de Huawei pertenecen a Huawei Technologies Co., Ltd.
Todas las demás marcas registradas y los otros nombres comerciales mencionados en este documento son propiedad de sus respectivos titulares.

## Aviso

Es posible que la totalidad o parte de los productos, las funcionalidades y/o los servicios que figuran en el presente documento no se encuentren dentro del alcance de un contrato vigente entre Huawei Cloud y el cliente. Las funcionalidades, los productos y los servicios adquiridos se limitan a los estipulados en el respectivo contrato. A menos que un contrato especifique lo contrario, ninguna de las afirmaciones, informaciones ni recomendaciones contenidas en el presente documento constituye garantía alguna, ni expresa ni implícita.

Huawei está permanentemente preocupada por la calidad de los contenidos de este documento; sin embargo, ninguna declaración, información ni recomendación aquí contenida constituye garantía alguna, ni expresa ni implícita. La información contenida en este documento se encuentra sujeta a cambios sin previo aviso.

# Índice



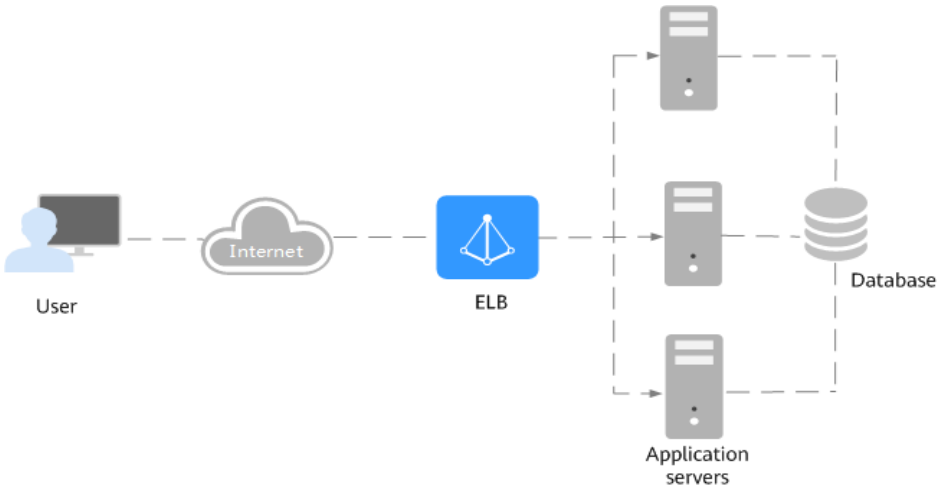
  

# 1 Qué es ELB

Elastic Load Balance (ELB) distribuye automáticamente el tráfico entrante entre varios servidores backend según las reglas de escucha que configure. ELB amplía las capacidades de servicio de sus aplicaciones y mejora su disponibilidad al eliminar los puntos únicos de fallo (SPOF).

Como se muestra en el ejemplo de la siguiente figura, ELB distribuye el tráfico entrante a tres servidores de aplicaciones, y cada servidor procesa un tercio de las solicitudes. ELB también proporciona comprobaciones de estado, que pueden detectar servidores no saludables. El tráfico se distribuye solo a los servidores que se ejecutan normalmente, lo que mejora la disponibilidad de las aplicaciones.

**Figura 1-1** Uso de un balanceador de carga




## Componentes de ELB

ELB consta de los siguientes componentes:

- **Balanceador de carga**: distribuye el tráfico entrante entre servidores backend en una o más zonas de disponibilidad (AZ).
- **Oyente**: utiliza el protocolo y el puerto que especifique para comprobar las solicitudes de los clientes y enrutar las solicitudes a los servidores backend asociados en función de las reglas de escucha y las políticas de reenvío que configure. Puede agregar uno o más oyentes a un balanceador de carga.

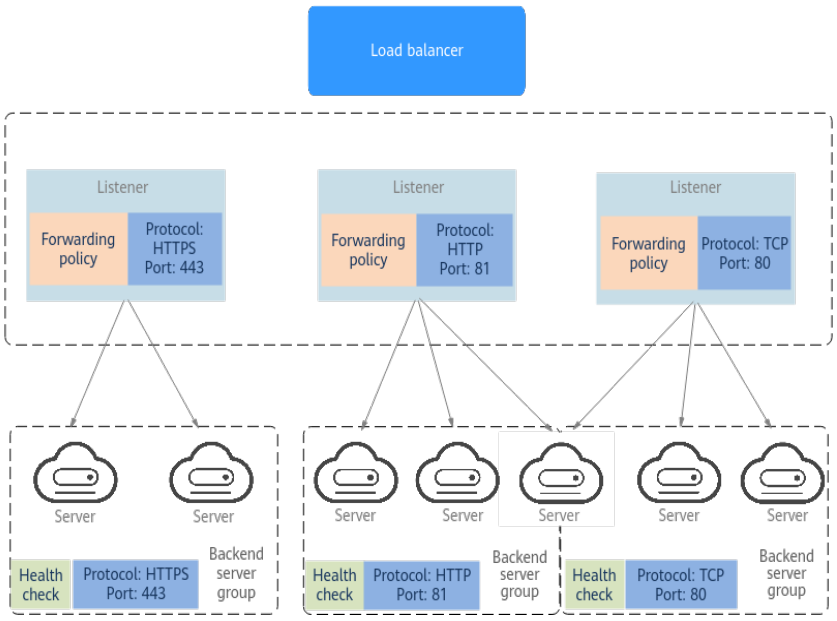
  

- **Grupo de servidores backend**: contiene uno o más servidores backend para recibir solicitudes enrutadas por el oyente. Debe agregar al menos un servidor backend a un grupo de servidores backend.

  Puede establecer una ponderación para cada servidor backend en función de su rendimiento.

  También puede configurar las comprobaciones de estado para un grupo de servidores backend para comprobar el estado de cada de backend. Cuando un servidor backend no está sano, el balanceador de carga deja de enrutar nuevas solicitudes a este servidor.

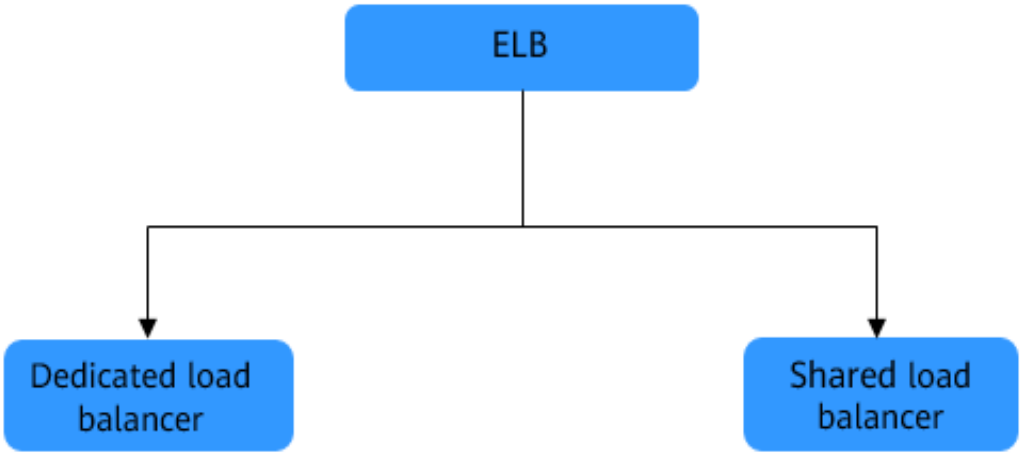
**Figura 1-2** Componentes de ELB




## Tipos de balanceadores de carga

ELB ofrece balanceadores de carga compartidos y balanceadores de carga dedicados.

**Figura 1-3** Tipos de balanceadores de carga



- Los balanceadores de carga dedicados tienen uso exclusivo de recursos subyacentes, de modo que el rendimiento de un balanceador de carga dedicado no se ve afectado por otros balanceadores de carga. Además, hay una amplia gama de especificaciones disponibles para la selección.

  NOTA

  Los balanceadores de carga dedicados no están disponibles en AF-Johannesburg y LA-Mexico City1.

- Los balanceadores de carga compartidos se despliegan en clústeres y comparten recursos subyacentes, de modo que el rendimiento de un balanceador de carga se ve afectado por otros balanceadores de carga. Los balanceadores de carga compartidos se denominaban anteriormente balanceadores de carga mejorados.

Para obtener más información sobre las diferencias entre los balanceadores de carga compartidos y dedicados, consulte **Tipos de productos de ELB**.

## Acceso a ELB

Puede utilizar cualquiera de los siguientes métodos para acceder a ELB:

- Consola de gestión

  Inicie sesión en la consola de gestión y seleccione **Elastic Load Balance (ELB)**.

- Las API

  Puede invocar a las API para acceder a ELB. Para obtener más información, consulta la ***Referencia de la API de Elastic Load Balance***.

# 2 Ventajas del producto

## Ventajas de los balanceadores de carga dedicados

- Sólido rendimiento

  Cada balanceador de carga tiene uso exclusivo de recursos aislados, cumpliendo con sus requisitos para manejar un gran número de solicitudes. Un único balanceador de carga desplegado en una AZ puede manejar hasta 20 millones de conexiones simultáneas.

  Si desplegar un balanceador de carga en varias AZ, su rendimiento, como el número de nuevas conexiones y el número de conexiones simultáneas se multiplicará. Por ejemplo, si despliega un balanceador de carga dedicado en dos AZ, puede manejar hasta 40 millones de conexiones simultáneas.

  NOTA

  - Si las solicitudes provienen de Internet, el balanceador de carga en cada AZ que seleccione encamina las solicitudes basadas en las direcciones IP de origen. Si desplegar un balanceador de carga en dos AZ, las solicitudes que los balanceadores de carga pueden manejar se duplicarán.
  - Para solicitudes de una red privada:
    - Si los clientes están en la AZ seleccionada al crear el balanceador de carga, las solicitudes son distribuidas por el balanceador de carga en esta AZ. Si el balanceador de carga no está disponible, las solicitudes son distribuidas por el balanceador de carga en otra AZ seleccionada.

      Si el balanceador de carga está disponible pero las conexiones que el balanceador de carga necesita manejar exceden la cantidad definida en las especificaciones, el servicio puede interrumpirse. Para solucionar este problema, necesita actualizar las especificaciones. Puede monitorear el uso del tráfico en la red privada por AZ.
    - Si los clientes están en una AZ que no está seleccionada al crear el balanceador de carga, el balanceador de carga distribuye las solicitudes en cada AZ que seleccione en función de las direcciones IP de origen.
  - Si las solicitudes provienen de una conexión de Direct Connect, el balanceador de carga de la misma AZ que la conexión de Direct Connect enruta las solicitudes. Si el balanceador de carga en esta AZ no está disponible, las solicitudes son distribuidas por el balanceador de carga en otra AZ.
  - Si los clientes están en una VPC que es diferente de donde funciona el balanceador de carga, el balanceador de carga en la AZ donde reside la subred de VPC original enruta las solicitudes. Si el balanceador de carga en esta AZ no está disponible, las solicitudes son distribuidas por el balanceador de carga en otra AZ.
- Disponibilidad alta

ELB puede enrutar el tráfico de forma ininterrumpida. Si sus servidores en una AZ no están sanos, automáticamente encamina el tráfico a servidores sanos en otras AZ. ELB proporciona un sistema integral de comprobación de estado para garantizar que el tráfico entrante se enrute solo a servidores backend sanos, mejorando la disponibilidad de sus aplicaciones.

- Seguridad ultraalta

  ELB soporta TLS 1.3 y puede enrutar solicitudes de HTTPS a servidores backend. Puede seleccionar políticas de seguridad o personalizar las políticas de seguridad que se ajusten a sus requisitos de seguridad.

- Múltiples protocolos

  ELB admite Quick UDP Internet Connection (QUIC), TCP, UDP, HTTP y HTTPS, para que puedan enrutar solicitudes a diferentes tipos de aplicaciones.

- Flexibilidad alta

  ELB puede enrutar solicitudes en función de su contenido, como el método de solicitud, el encabezado, el URL, la ruta y la dirección IP de origen. También pueden redirigir solicitudes a otro oyente o URL, o devolver una respuesta fija a los clientes.

- Sin límite

  ELB puede enrutar solicitudes tanto a servidores en la nube como en las instalaciones, lo que le permite aprovechar los recursos de la nube para manejar el tráfico de ráfagas.

- Facilidad de uso

  ELB proporciona un conjunto diverso de algoritmos que le permiten configurar diferentes políticas de enrutamiento de tráfico para satisfacer sus requisitos, al tiempo que mantiene los despliegues simples.

## Ventajas de los balanceadores de carga compartidos

- Rendimiento alto

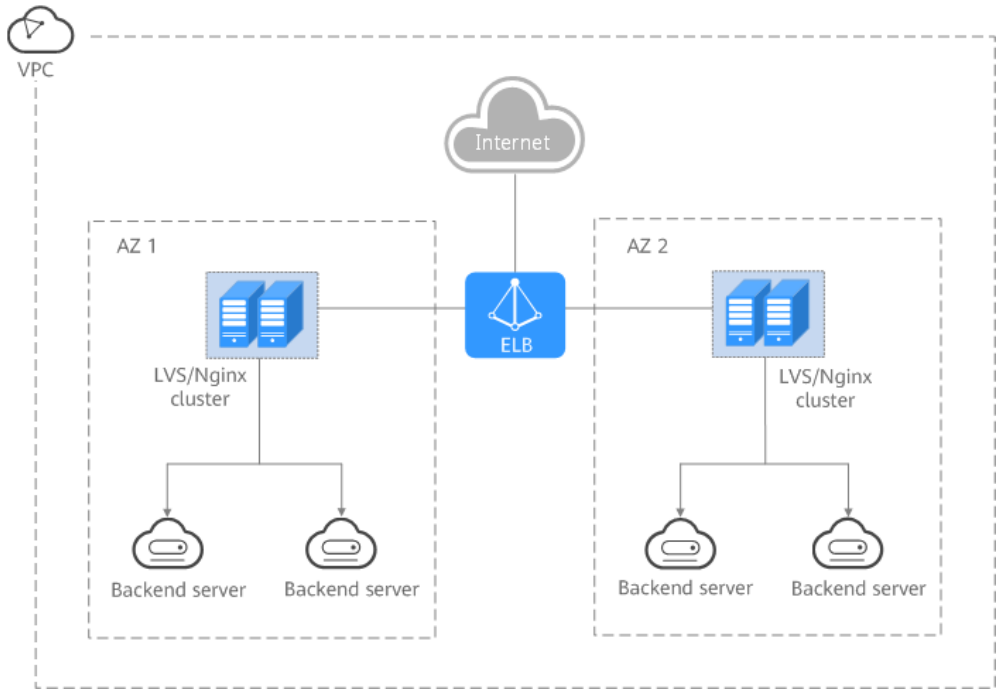
  Los balanceadores de carga compartidos proporcionan un rendimiento garantizado, que puede manejar hasta 50,000 conexiones simultáneas, 5,000 nuevas conexiones por segundo y 5,000 consultas por segundo.

- Disponibilidad alta

  Los balanceadores de carga de Compartido pueden enrutar el tráfico a través de AZ, asegurando que sus servicios sean ininterrumpidos. Si los servidores de una AZ no están sanos, ELB enruta automáticamente el tráfico a servidores sanos en otras AZ. Los balanceadores de carga Compartidos proporcionan un mecanismo completo de comprobación de estado para garantizar que el tráfico entrante se enrute solo a servidores backend sanos, mejorando la disponibilidad de sus aplicaciones.

**Figura 2-1** Disponibilidad alta




- Múltiples protocolos

  ELB admite protocolos TCP, UDP, HTTP y HTTPS para enrutar solicitudes a diferentes tipos de aplicaciones.
- Facilidad de uso

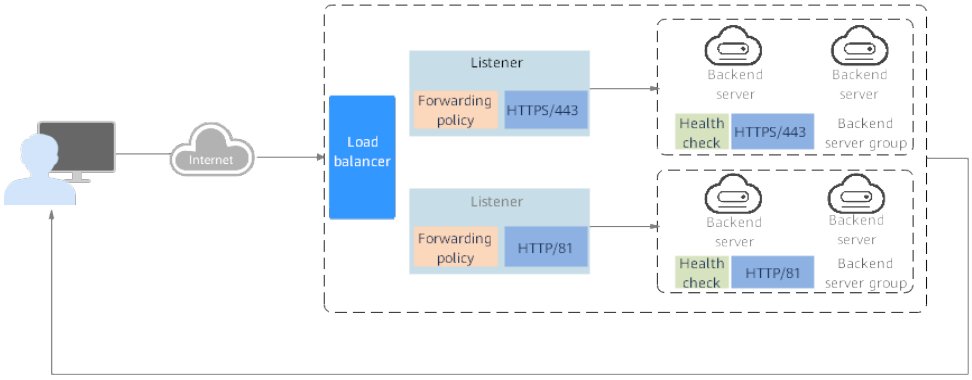
  ELB proporciona un conjunto diverso de algoritmos que le permiten configurar diferentes políticas de enrutamiento de tráfico para satisfacer sus requisitos, al tiempo que mantiene los despliegues simples.
- Alto nivel de confiabilidad

  Los balanceadores de carga se despliegan en dos AZ y pueden distribuir el tráfico de manera más uniforme.

# 3 Cómo funciona ELB

**Figura 3-1** Cómo funciona ELB




A continuación se describe cómo funciona ELB:

1. Un cliente envía una solicitud a su aplicación.
2. Los oyentes agregados a su balanceador de carga utilizan los protocolos y puertos que ha configurado para recibir la solicitud.
3. El oyente reenvía la solicitud al grupo de servidores backend asociado en función de su configuración. Si ha configurado una política de reenvío para el oyente, el oyente evalúa la solicitud basándose en la política de reenvío. Si la solicitud coincide con la política de reenvío, el oyente reenvía la solicitud al grupo de servidores backend configurado para la política de reenvío.
4. Los servidores backend sanos del grupo de servidores backend reciben la solicitud basada en el algoritmo de equilibrio de carga y las reglas de enrutamiento especificadas en la política de reenvío, gestionan la solicitud y devuelven un resultado al cliente.

La forma en que se enrutan las solicitudes depende de los **algoritmos de equilibrio de carga** configurados para cada grupo de servidores backend. Si el oyente utiliza HTTP o HTTPS, la forma en que se enrutan las solicitudes también depende de **las políticas de reenvío** configuradas para el oyente.

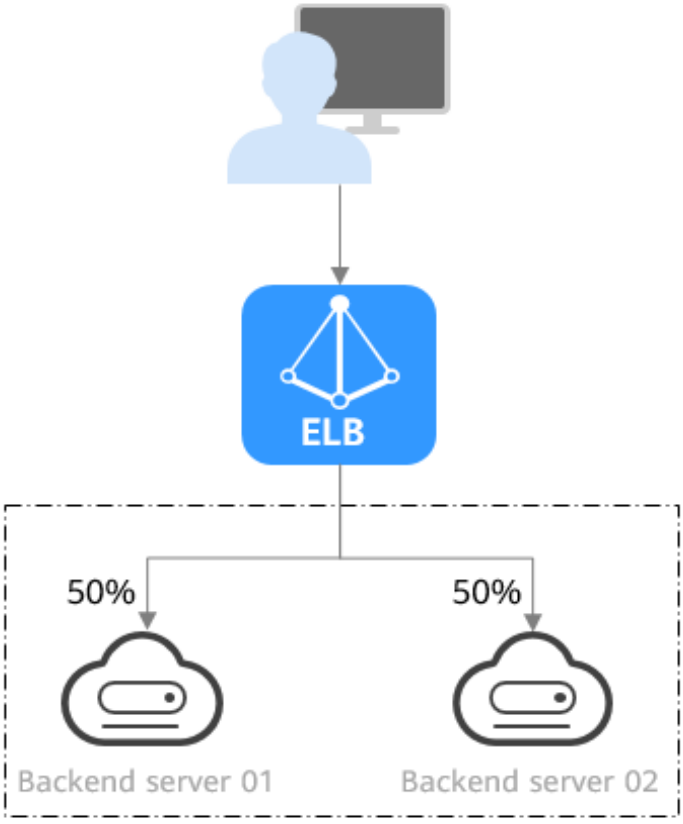
## Algoritmos de equilibrio de carga

Los balanceadores de carga dedicados admiten cuatro algoritmos de balanceo de carga: round robin ponderado, conexiones mínimas ponderadas, hash IP de origen e ID de conexión. Los balanceadores de carga compartidos admiten tres algoritmos de balanceo de carga: round robin ponderado, conexiones mínimas ponderadas y hash IP de origen.

- Round robin ponderado: las solicitudes se enrutan a los servidores backend utilizando el algoritmo round robin. Los servidores back-end con mayores pesos reciben proporcionalmente más solicitudes, mientras que los servidores con igual peso reciben el mismo número de solicitudes. Este algoritmo se utiliza a menudo para conexiones cortas, como conexiones HTTP.

  La siguiente figura muestra un ejemplo de cómo se distribuyen las solicitudes usando el algoritmo round robin ponderado. Dos servidores backend están en la misma AZ y tienen el mismo peso, y cada servidor recibe la misma proporción de solicitudes.

  **Figura 3-2** Distribución del tráfico utilizando el algoritmo de round robin ponderado



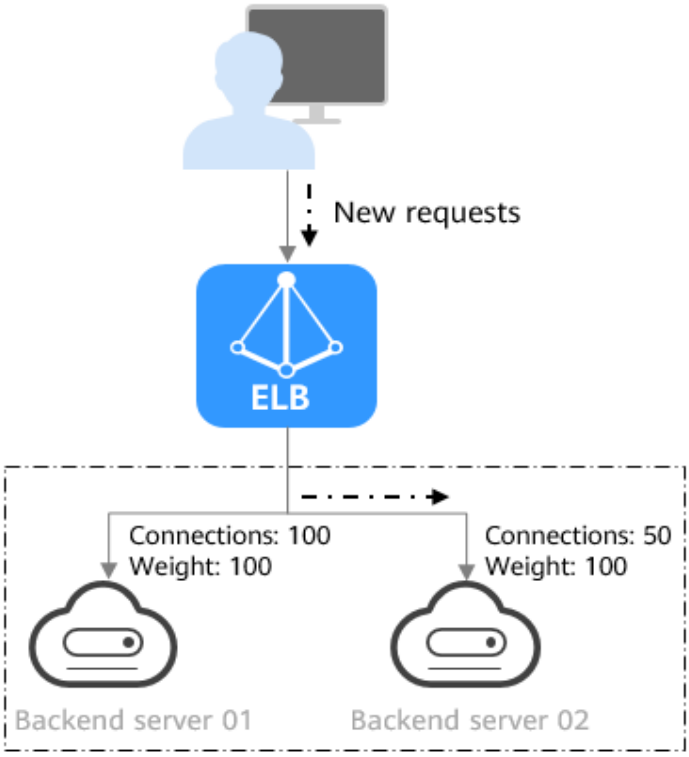

- Conexiones mínimas ponderadas: Además del peso asignado a cada servidor, también se tiene en cuenta el número de conexiones procesadas por cada servidor backend. Las solicitudes se enrutan al servidor con la relación de conexiones a ponderación más baja. Además del número de conexiones, a cada servidor se le asigna una ponderación basada en su capacidad. Las solicitudes se enrutan al servidor con la relación de conexiones a ponderación más baja. Este algoritmo se utiliza a menudo para conexiones persistentes, como conexiones a una base de datos.

  La siguiente figura muestra un ejemplo de cómo se distribuyen las solicitudes usando el algoritmo de conexiones mínimas ponderadas. Dos servidores de backend están en la

misma AZ y tienen la misma ponderación, se han establecido 100 conexiones con el servidor backend 01, y se han conectado 50 conexiones con el servidor backend 02. Las nuevas peticiones se encaminan preferentemente al servidor de backend 02.

**Figura 3-3** Distribución del tráfico utilizando el algoritmo de conexiones mínimas ponderadas

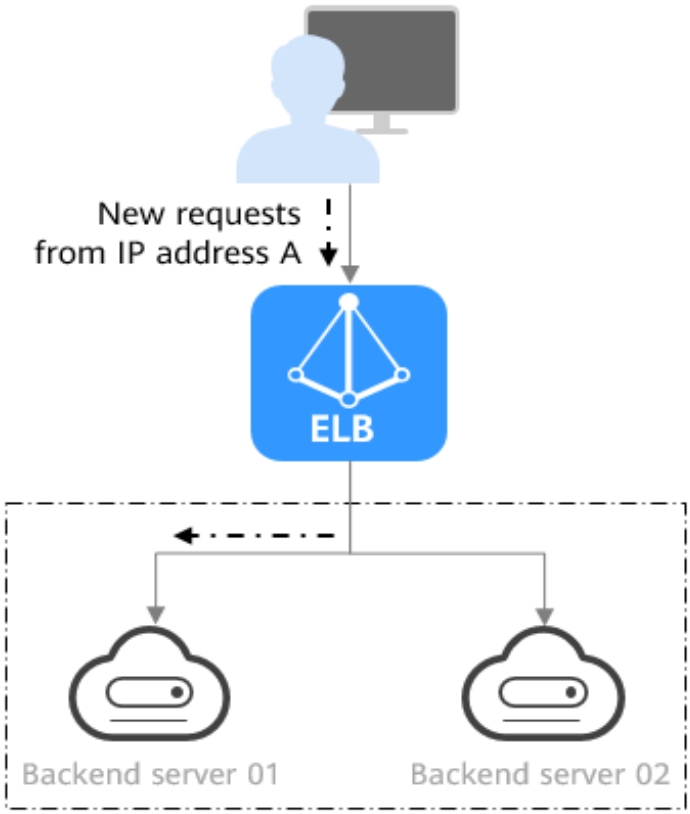



- Hash IP de origen: La dirección IP de origen de cada solicitud se calcula utilizando el algoritmo de hash consistente para obtener una clave de hash única, y todos los servidores backend están numerados. La clave generada se utiliza para asignar el cliente a un servidor en particular. Esto permite que las solicitudes de diferentes clientes se enruten en función de las direcciones IP de origen y garantiza que un cliente se dirija al mismo servidor que estaba usando anteriormente. Este algoritmo funciona bien para conexiones de TCP de balanceadores de carga que no usan cookies.

  La siguiente figura muestra un ejemplo de cómo se distribuyen las solicitudes utilizando el algoritmo hash IP de origen. Dos servidores backend están en la misma AZ y tienen la misma ponderación. Si el servidor backend 01 ha procesado una solicitud desde la dirección IP A, el balanceador de carga encaminará nuevas solicitudes desde la dirección IP A al servidor backend 01.

**Figura 3-4** Distribución del tráfico mediante el algoritmo de hash IP de origen




- ID de conexión: El ID de conexión en el paquete se calcula utilizando el algoritmo hash consistente para obtener un valor específico, y los servidores backend se numeran. El valor generado determina a qué servidor backend se enrutan las solicitudes. Esto permite que las solicitudes con diferentes ID de conexión se enruten a diferentes servidores backend y garantiza que las solicitudes con el mismo ID de conexión se enruten al mismo servidor backend. Este algoritmo se aplica a las solicitudes de QUIC.

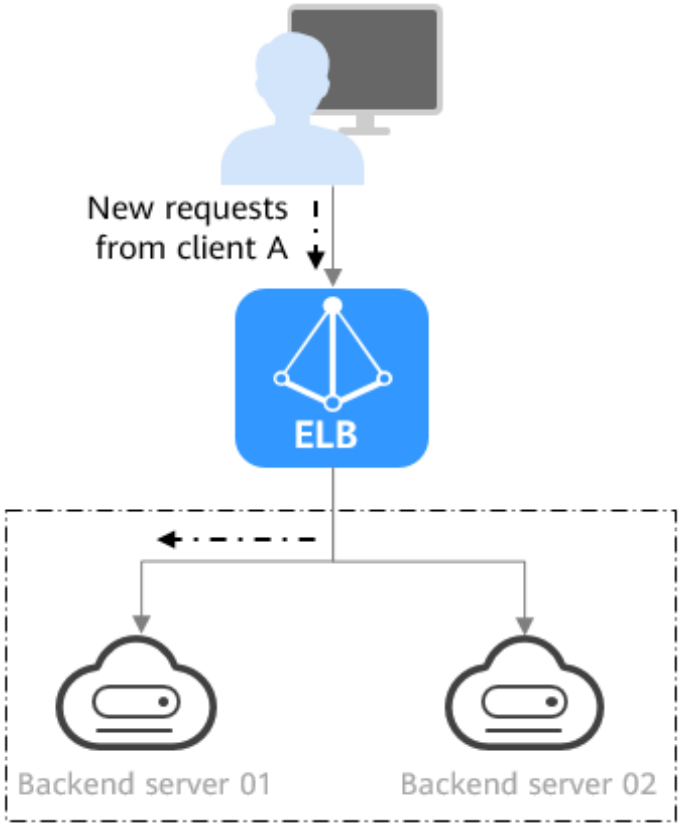
  NOTA

  Actualmente, solo los balanceadores de carga dedicados admiten el algoritmo de ID de conexión.

  **Figura 3-5** muestra un ejemplo de cómo se distribuyen las solicitudes usando el algoritmo de ID de conexión. Dos servidores backend están en la misma AZ y tienen la misma ponderación. Si el servidor de backend 01 ha procesado una solicitud del cliente A, el balanceador de carga encaminará nuevas solicitudes desde el cliente A al servidor backend 01.

**Figura 3-5** Distribución del tráfico mediante el algoritmo de ID de conexión




## Factores que afectan el equilibrio de carga

Además del algoritmo de equilibrio de carga, los factores que afectan al equilibrio de carga generalmente incluyen el tipo de conexión, la adherencia de sesión y los pesos del servidor.

Suponga que hay dos servidores backend con la misma ponderación (no cero), se selecciona el algoritmo de conexiones mínimas ponderadas, las sesiones adhesivas no están habilitadas, y se han establecido 100 conexiones con el servidor backend 01, y 50 conexiones con el servidor backend 02.

Cuando el cliente A desea acceder al servidor backend 01, el balanceador de carga establece una conexión persistente con el servidor backend 01 y encamina continuamente las solicitudes desde el cliente A al servidor backend 01 antes de que se desconecte la conexión persistente. Cuando otros clientes acceden a servidores backend, el balanceador de carga encamina las solicitudes al servidor backend 02 usando el algoritmo de conexiones mínimas ponderadas.

> **NOTA**
>
> Si los servidores backend se declaran no sanos o sus ponderaciones se establecen en 0, el balanceador de carga no enrutará ninguna solicitud a los servidores backend.

Para obtener detalles sobre el algoritmo de menos conexiones ponderadas, consulte **Algoritmos de equilibrio de carga**.

Si las solicitudes no se enrutan uniformemente, solucione el problema realizando las operaciones descritas en el documento **¿Cómo puedo comprobar si el tráfico está distribuido uniformemente?**

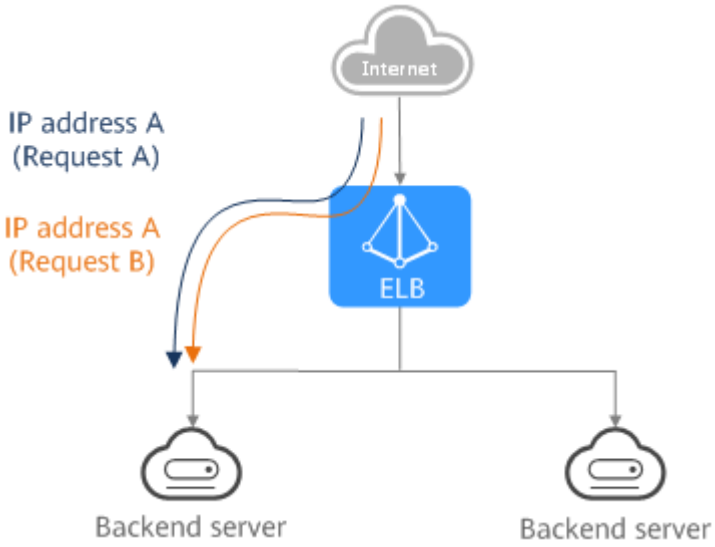
# 4 Escenarios de la aplicación

## Aplicaciones de tráfico pesado

Para una aplicación con mucho tráfico, como un gran portal o una tienda de aplicaciones móviles, ELB distribuye uniformemente el tráfico entrante entre los servidores backend, equilibrando la carga y asegurando un rendimiento constante.

Las sesiones persistentes garantizan que las solicitudes de un cliente se reenvíen siempre al mismo servidor backend para un procesamiento rápido.

**Figura 4-1** Sesión persistente

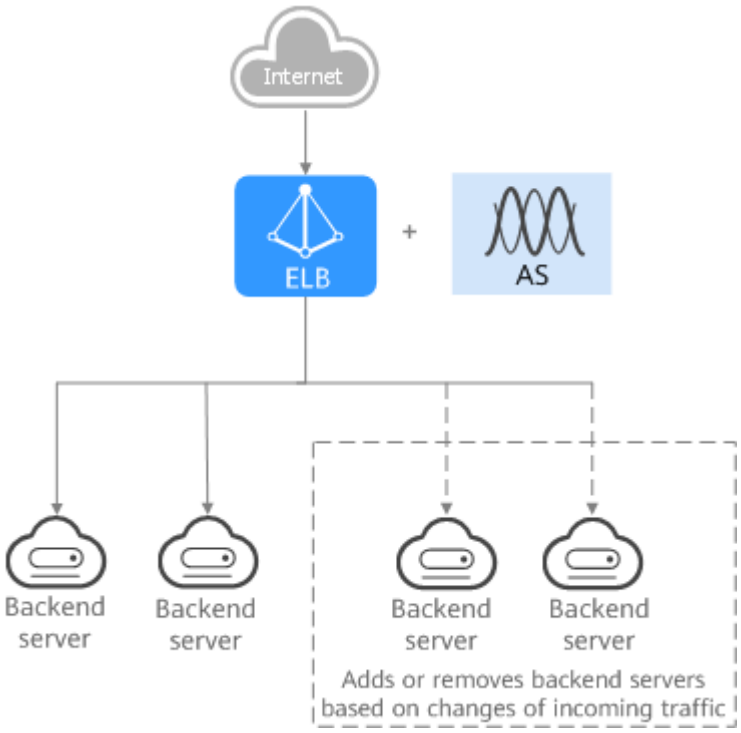



## Aplicaciones con picos y canales predecibles en el tráfico

Para una aplicación que tiene picos y valles predecibles en los volúmenes de tráfico, ELB trabaja con Auto Scaling para agregar servidores automáticamente durante las promociones cuando hay picos de tráfico repentinos y luego eliminarlos cuando el tráfico vuelve a la normalidad. Esto le ayuda a mejorar la disponibilidad de recursos y reducir los costos de TI.

**Figura 4-2** Escalabilidad flexible



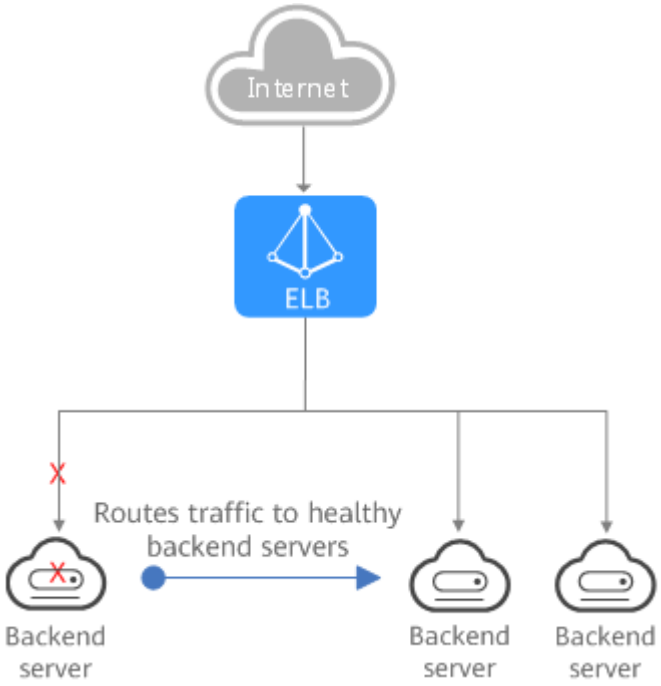

## Cero SPOF

ELB realiza rutinariamente comprobaciones de estado en los servidores backend para supervisar su estado. Si se detecta un servidor de back-end que no está en buen estado, ELB no encaminará las solicitudes a este servidor hasta que se recupere.

Esto hace que ELB sea una buena opción para ejecutar servicios que requieren una alta fiabilidad, como sitios web y sistemas de cobro de peajes.

**Figura 4-3** Eliminación de los SPOF

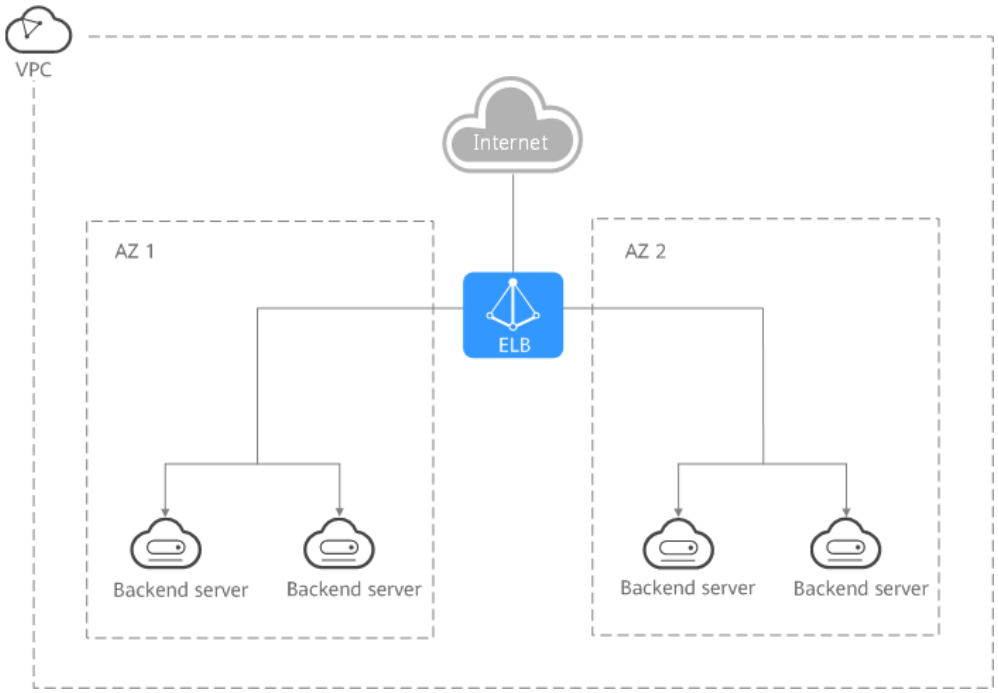



## Balanceo de carga inter-AZ

ELB puede distribuir el tráfico entre las zonas de disponibilidad. Cuando una AZ se vuelve defectuosa, ELB distribuye el tráfico a través de servidores backend en otras AZ.

ELB es ideal para la banca, la vigilancia y los sistemas de aplicaciones de gran tamaño que requieren alta disponibilidad.

**Figura 4-4** Distribución de tráfico a servidores en una o más AZ



# 5 Differences Between Dedicated and Shared Load Balancers

## 5.1 Tipos de productos de ELB

### Introducción a ELB

Elastic Load Balance (ELB) distribuye automáticamente el tráfico entrante entre servidores para equilibrar sus cargas de trabajo, lo que aumenta las capacidades del servicio y la tolerancia a fallas de las aplicaciones del usuario. ELB amplía las capacidades de servicio de sus aplicaciones.

### Tipos de balanceadores de carga

ELB proporciona balanceadores de carga compartidos y balanceadores de carga dedicados para que usted elija.

**Tabla 5-1** Tipos de balanceadores de carga

| Concepto | Balanceador de carga dedicado | Balanceador de carga compartido |
| --- | --- | --- |
| Modo de despliegue | Usted obtiene acceso exclusivo a los recursos del balanceador de carga. El rendimiento de un balanceador de carga dedicado nunca se ve afectado por las cargas en otros balanceadores de carga. Además, hay una amplia gama de especificaciones disponibles para que usted elija. | Se despliegan en clústeres y comparten recursos con otras instancias. Soportan un rendimiento garantizado. |

| Concepto | Balanceador de carga dedicado | Balanceador de carga compartido |
| --- | --- | --- |
| Especificaciones | ● Especificaciones elásticas: se le cobra por el tiempo que se está ejecutando cada balanceador de carga y el número de LCU que utiliza.<br>● Especificaciones fijas: Varias especificaciones están disponibles para que usted seleccione para satisfacer mejor sus necesidades.<br>Para obtener más información, véase **Especificaciones de los balanceadores de carga dedicados**. | - |
| Rendimiento | Un balanceador de carga dedicado en una AZ puede establecer hasta **20 millones de conexiones simultáneas**. Si desplegar un balanceador de carga dedicado en dos AZ, el número de conexiones simultáneas se duplicará.<br>Por ejemplo, si despliega un balanceador de carga dedicado en dos AZ, puede manejar hasta 40 millones de conexiones simultáneas. | Si está habilitado el rendimiento garantizado, los balanceadores de carga compartidos pueden manejar hasta 50,000 conexiones simultáneas, 5,000 nuevas conexiones por segundo y 5,000 consultas por segundo. |

| Concepto | Balanceador de carga dedicado | Balanceador de carga compartido |
| --- | --- | --- |
| AZ | Puede seleccionar una o más AZ según sea necesario.<br>● Si las solicitudes provienen de Internet, el balanceador de carga en cada AZ que seleccione encamina las solicitudes basadas en las direcciones IP de origen. Si desplegar un balanceador de carga en dos AZ, las solicitudes que los balanceadores de carga pueden manejar se duplicarán.<br>● Para solicitudes de una red privada:<br>– Si los clientes están en la AZ seleccionada al crear el balanceador de carga, las solicitudes son distribuidas por el balanceador de carga en esta AZ. Si el balanceador de carga no está sano, las solicitudes son distribuidas por el balanceador de carga en otra AZ seleccionada.<br>Si el balanceador de carga está en buen estado pero las conexiones que el balanceador de carga necesita manejar exceden la cantidad definida en las especificaciones, el servicio puede interrumpirse. Para solucionar este problema, necesita actualizar las especificaciones. Puede monitorear el uso del tráfico en la red privada por AZ.<br>– Si los clientes están en una AZ que no está seleccionada al crear el balanceador de carga, el balanceador de carga distribuye las solicitudes en cada AZ que seleccione en función de las direcciones IP de origen. | - |

| Concepto | Balanceador de carga dedicado | Balanceador de carga compartido |
| --- | --- | --- |
| | ● Si las solicitudes provienen de una conexión de Direct Connect, el balanceador de carga de la misma AZ que la conexión de Direct Connect enruta las solicitudes. Si el balanceador de carga no está disponible, las solicitudes son distribuidas por el balanceador de carga en otra AZ.<br>● Si los clientes están en una VPC que es diferente de donde funciona el balanceador de carga, el balanceador de carga en la AZ donde reside la subred de VPC original enruta las solicitudes. Si el balanceador de carga no está disponible, las solicitudes son distribuidas por el balanceador de carga en otra AZ. | |
| Concepto de facturación | ● Especificaciones fijas: facturadas por las LCU según las especificaciones que seleccione.<br>● Especificaciones elásticas: facturadas por la cantidad de LCU que usa y por cuánto tiempo usa sus balanceadores de carga | Se le cobrará por el tiempo que utilice cada balanceador de carga si está habilitado el rendimiento garantizado. |

## Comparación de funciones

**Tabla 5-2** Comparación de funciones

| Concepto | Balanceador de carga dedicado | Balanceador de carga compartido |
| --- | --- | --- |
| Capacidades | Potentes capacidades para procesar solicitudes de Capa 4 y Capa 7, políticas avanzadas de reenvío y múltiples protocolos. | Capacidades básicas para procesar solicitudes de Capa 4 y Capa 7 |

| Concepto | Balanceador de carga dedicado | Balanceador de carga compartido |
| --- | --- | --- |
| Escenarios de aplicación | Servicios de alto tráfico y altamente simultáneos, como sitios web grandes, aplicaciones nativas de la nube, IoV y aplicaciones de recuperación ante desastres multi-AZ | Servicios con poco tráfico, como sitios web pequeños y aplicaciones de HA comunes |
| **Protocolos frontend** | TCP, UDP, HTTP y HTTPS | TCP, UDP, HTTP y HTTPS |
| **Protocolos backend** | TCP, UDP, HTTP, HTTPS y QUIC | TCP, UDP y HTTP |
| **Capacidades de reenvío** | Proporcione potentes capacidades de procesamiento de Capa 4 y Capa 7 para **reenviar solicitudes** basadas en lo siguiente:<br>● Reglas de reenvío: nombre de dominio, URL, método de solicitud de HTTP, encabezado de HTTP, cadena de consulta y bloque CIDR<br>● Acciones: reenviar a un grupo de servidores backend, redirigir a otro oyente, redirigir a otro URL, reescribir y devolver un cuerpo de respuesta específico | Proporcione capacidades básicas de procesamiento de Capa 4 y Capa 7 para **reenviar solicitudes** basadas en lo siguiente:<br>● Reglas de reenvío: nombre de dominio y URL<br>● Acciones: reenviar a un grupo de servidores backend y redirigir a otro oyente |
| **Funciones clave de los grupos de servidores backend** | ● Comprobación de estado<br>● Sesión adhesiva<br>● Inicio lento | ● Comprobación de estado<br>● Sesión adhesiva |
| **Algoritmos de equilibrio de carga** | ● Round robin ponderado<br>● Conexiones mínimas ponderadas<br>● Hash de IP de origen<br>● ID de conexión | ● Round robin ponderado<br>● Conexiones mínimas ponderadas<br>● Hash de IP de origen |
| **Modos de reenvío de grupos de servidores backend** | ● Equilibrio de carga<br>● Activo/en espera | Equilibrio de carga |

| Concepto | Balanceador de carga dedicado | Balanceador de carga compartido |
| --- | --- | --- |
| Tipo de backend | • ECS<br>• IP como servidor backend<br>• Interfaz de red suplementaria<br>• BMS<br>• Clúster de Turbo de CCE | • ECS<br>• BMS<br>• Clúster de Turbo de CCE |

# 5.2 Detalles de comparación de características

## Protocolos

**Tabla 5-3** Protocolos soportados por cada tipo de balanceador de carga

| Protocolo | Descripción | Balanceador de carga dedicado | Balanceador de carga compartido |
| --- | --- | --- | --- |
| TCP/UDP (Capa 4) | Después de recibir solicitudes de TCP o de UDP de los clientes, el balanceador de carga enruta directamente las solicitudes a los servidores backend.<br>El equilibrio de carga en la capa 4 presenta una alta eficiencia de enrutamiento. | √ | √ |
| HTTP/HTTPS (Capa 7) | Después de recibir una solicitud de acceso, el oyente necesita identificar la solicitud y reenviar datos basándose en los campos en la cabecera del paquete HTTP/HTTPS.<br>El equilibrio de carga en la capa 7 proporciona algunas características avanzadas, como la transmisión cifrada y las sesiones adhesivas basadas en cookies. | √ | √ |
| Compatibilidad con HTTPS | HTTPS se puede utilizar como protocolo frontend y backend. | √ | x |

| Protocolo | Descripción | Balanceador de carga dedicado | Balanceador de carga compartido |
|---|---|---|---|
| QUIC | Si utiliza UDP y QUIC como protocolo frontend, puede seleccionar QUIC como protocolo backend y seleccionar el algoritmo de ID de conexión para enrutar solicitudes con el mismo ID de conexión al mismo servidor backend.<br>QUIC tiene las ventajas de baja latencia, alta confiabilidad y sin bloqueo de cabecera (bloqueo HOL), y es muy adecuado para Internet móvil. No es necesario establecer nuevas conexiones cuando se cambia entre una red Wi-Fi y una red móvil. | √ | x |
| HTTP/2 | Hypertext Transfer Protocol 2.0 (HTTP/2) es una nueva versión del protocolo HTTP. Es compatible con HTTP/1.X y proporciona un rendimiento y seguridad mejorados.<br>Solo los oyentes de HTTPS admiten esta característica. | √ | √ |
| WebSocket | WebSocket es un nuevo protocolo de HTML5 que proporciona comunicación full duplex entre el navegador y el servidor. WebSocket ahorra recursos del servidor y ancho de banda, y permite la comunicación en tiempo real. | √ | √ |

## Configuraciones de red

**Tabla 5-4** Comparación de configuración de red

| Función | Descripción | Balanceador de carga dedicado | Balanceador de carga compartido |
|---|---|---|---|
| Red IPv4 pública | El balanceador de carga enruta las solicitudes de los clientes a los servidores backend por Internet. | √ | √ |
| Red IPv4 privada | El balanceador de carga enruta las solicitudes de los clientes a los servidores backend en una VPC. | √ | √ |

| Función | Descripción | Balanceador de carga dedicado | Balanceador de carga compartido |
|---|---|---|---|
| Red IPv6 | Los balanceadores de carga pueden enrutar solicitudes de clientes IPv6. | √ | x |
| Cambio de una dirección IPv4 privada | Puede cambiar la dirección IPv4 privada a otra en la subred actual u otras subredes. | √ | x |
| Vinculación o desvinculación de una EIP | Puede vincular una EIP a un balanceador de carga o desvincular la EIP de un balanceador de carga según los requisitos de servicio. | √ | √ |
| Modificación del ancho de banda | Puede cambiar el ancho de banda de los balanceadores de carga de red pública según sea necesario. | √ | √ |

## Características clave de oyentes

**Tabla 5-5** Comparación de características clave

| Función | Descripción | Balanceador de carga dedicado | Balanceador de carga compartido |
|---|---|---|---|
| Access Control | Puede agregar direcciones IP a una lista blanca o negra para controlar el acceso a un oyente.<br>● Una lista blanca permite que las direcciones IP especificadas accedan al listener.<br>● Una lista negra deniega el acceso desde direcciones IP especificadas. | √ | √ |
| Mutual Authentication | Esta característica permite que los clientes y el balanceador de carga se autentican entre sí. Solo los clientes autenticados podrán acceder al balanceador de carga.<br>La autenticación mutua solo es compatible con los oyentes de HTTPS. | √ | √ |

| Función | Descripción | Balanceador de carga dedicado | Balanceador de carga compartido |
|---|---|---|---|
| SNI | La indicación de nombre de servidor (SNI) es una extensión de TLS y se utiliza cuando un servidor utiliza varios nombres de dominio y certificados. Una vez habilitado el SNI, se requieren los certificados correspondientes a los nombres de dominio.<br>SNI solo se puede habilitar para oyentes de HTTPS. | √ | √ |
| Transfer Client IP Address | Esta característica permite a los servidores backend obtener las direcciones IP reales de los clientes.<br>Esta función está habilitada para balanceadores de carga dedicados de forma predeterminada y no se puede deshabilitar. | √ | √ |
| Características avanzadas de los oyentes de HTTP/HTTPS | | | |
| Default Security Policy | Le permite seleccionar las políticas de seguridad adecuadas para mejorar la seguridad del servicio al agregar oyentes de HTTPS. Una política de seguridad es una combinación de protocolos TLS y conjuntos de cifrado. | √ | √ |
| Custom Security Policy | Le permite seleccionar un protocolo TLS y un conjunto de cifrado para personalizar una política de seguridad al agregar oyentes de HTTPS. | √ | x |
| Transfer Load Balancer EIP | Permite almacenar la EIP vinculada al balanceador de carga en el encabezado X-Forwarded-ELB-IP y pasarlo a los servidores backend. | √ | √ |

## Capacidades de reenvío

Puede agregar políticas de reenvío a oyentes de HTTP o de HTTPS para reenviar solicitudes a diferentes grupos de servidores backend. Las políticas de reenvío avanzadas solo están disponibles para balanceadores de carga dedicados.

Puede establecer reglas y acciones de reenvío para una política de reenvío. Para más detalles, véase **Tabla 5-6** y **Tabla 5-7**.

**Tabla 5-6** Reglas de reenvío compatibles con cada tipo de balanceador de carga

| Regla de reenvío | Descripción | Balanceador de carga dedicado | Balanceador de carga compartido |
|---|---|---|---|
| Nombre de dominio | Enrutar solicitudes basadas en el nombre de dominio. El nombre de dominio de la solicitud debe coincidir exactamente con el de la política de reenvío. | √ | √ |
| URL | Enrutar solicitudes basadas en las URL.<br>Hay tres reglas de coincidencia de URL: coincidencia exacta, coincidencia de prefijo y coincidencia de expresión regular. | √ | √ |
| Método de solicitud de HTTP | Enrutar solicitudes basadas en el método de HTTP.<br>Las opciones incluyen GET, POST, PUT, DELETE, PATCH, HEAD y OPTIONS. | √ | x |
| Encabezado de HTTP | Enrutar solicitudes basadas en el encabezado de HTTP.<br>Un encabezado HTTP consta de una clave y uno o más valores. Es necesario configurar la clave y los valores por separado. | √ | x |
| Query string | Enrutar solicitudes basadas en la cadena de consulta. | √ | x |
| Bloque CIDR | Enrutar las solicitudes basadas en las direcciones IP de origen desde donde se originan las solicitudes. | √ | x |

**Tabla 5-7** Acciones admitidas por cada tipo de balanceador de carga

| Acción | Descripción | Balanceador de carga dedicado | Balanceador de carga compartido |
|---|---|---|---|
| Reenvío a un grupo de servidores backend | Reenviar solicitudes al grupo de servidores backend especificado. | √ | √ |

| Acción | Descripción | Balanceador de carga dedicado | Balanceador de carga compartido |
|---|---|---|---|
| Redirigir a otro oyente | Redirigir las solicitudes a un oyente de HTTPS, que luego enruta las solicitudes a su grupo de servidores backend asociado. | √ | x |
| Redirigir a otro URL | Redirigir las solicitudes al URL configurado.<br>Cuando los clientes acceden al sitio web A, el balanceador de carga devuelve 302 o cualquier otro código de estado 3xx y redirige automáticamente a los clientes al sitio web B. Puede personalizar el URL de redirección que se devolverá a los clientes. | √ | x |
| Devolver un cuerpo de respuesta específico | Devolver una respuesta fija a los clientes.<br>Puede personalizar el código de estado y el cuerpo de respuesta que los balanceadores de carga devuelven directamente a los clientes sin necesidad de enrutar las solicitudes a los servidores backend. | √ | x |

## Características clave de los grupos de servidores backend

**Tabla 5-8** Características clave compatibles con cada tipo de balanceador de carga

| Característica clave | Descripción | Balanceador de carga dedicado | Balanceador de carga compartido |
|---|---|---|---|
| Comprobación de estado | ELB envía periódicamente solicitudes a los servidores backend para comprobar sus estados de ejecución. Este proceso se llama comprobación de estado. Puede realizar comprobaciones de estado para determinar si hay un servidor backend disponible. | √ | √ |
| Sesión adhesiva | Las solicitudes del mismo cliente se encaminarán al mismo servidor backend durante la sesión. | √ | √ |

| Característica clave | Descripción | Balanceador de carga dedicado | Balanceador de carga compartido |
|---|---|---|---|
| Inicio lento | El balanceador de carga aumenta linealmente la proporción de solicitudes a los nuevos servidores backend agregados al grupo de servidores backend.<br>El arranque lento da tiempo a las aplicaciones para calentar y responder a las solicitudes con un rendimiento óptimo. | √ | x |
| Reenvío activo/en espera | El balanceador de carga enruta el tráfico al servidor activo si funciona normalmente y al servidor en espera si el servidor activo no funciona correctamente.<br>Debe agregar dos servidores backend al grupo de servidores backend, uno que actúe como servidor activo y el otro como servidor en espera. | √ | x |

## Algoritmos de equilibrio de carga

**Tabla 5-9** Comparación del algoritmo de equilibrio de carga

| Algoritmo de balanceo de carga | Descripción | Balanceador de carga dedicado | Balanceador de carga compartido |
|---|---|---|---|
| Round robin ponderado | Enrutar las solicitudes a los servidores backend utilizando el algoritmo round robin. Los servidores backend con mayores pesos reciben proporcionalmente más solicitudes, mientras que los servidores con la misma ponderación reciben el mismo número de solicitudes. | √ | √ |
| Conexiones mínimas ponderadas | Enrutar solicitudes a servidores backend con la relación más pequeña (conexiones actuales divididas por ponderación). | √ | √ |
| Hash de IP de origen | Enrutar las solicitudes del mismo cliente al mismo servidor backend dentro de un período de tiempo. | √ | √ |

| Algoritmo de balanceo de carga | Descripción | Balanceador de carga dedicado | Balanceador de carga compartido |
|---|---|---|---|
| ID de conexión | Calcular la dirección IP de origen de cada solicitud usando el algoritmo de hash consistente para obtener una clave de hash única y enrutar las solicitudes al servidor particular basándose en la clave generada. | √ | x |

## Tipo de servidor backend

**Tabla 5-10** Tipos de servidor backend admitidos

| Tipo de servidor backend | Descripción | Balanceador de carga dedicado | Balanceador de carga compartido |
|---|---|---|---|
| IP como servidor backend | Puede agregar servidores en una VPC de otro extremo, en una VPC que esté en otra región y conectada con una conexión a la nube, o en un centro de datos local en el otro extremo de una conexión de Direct Connect o de VPN, mediante el uso de las direcciones IP del servidor. | √ | x |
| Interfaz de red suplementaria | Puede conectar interfaces de red suplementarias a servidores backend. | √ | x |
| ECS | Puede utilizar balanceadores de carga para distribuir el tráfico entrante entre los ECS. | √ | √ |
| BMS | Puede utilizar balanceadores de carga para distribuir el tráfico entrante entre los BMS. | √ | √ |
| Clúster de Turbo de CCE | Puede utilizar balanceadores de carga para distribuir el tráfico entrante entre los clústeres de Turbo de CCE. Para obtener más información, consulte la *Guía de usuario de Cloud Container Engine*. | √ | √ |

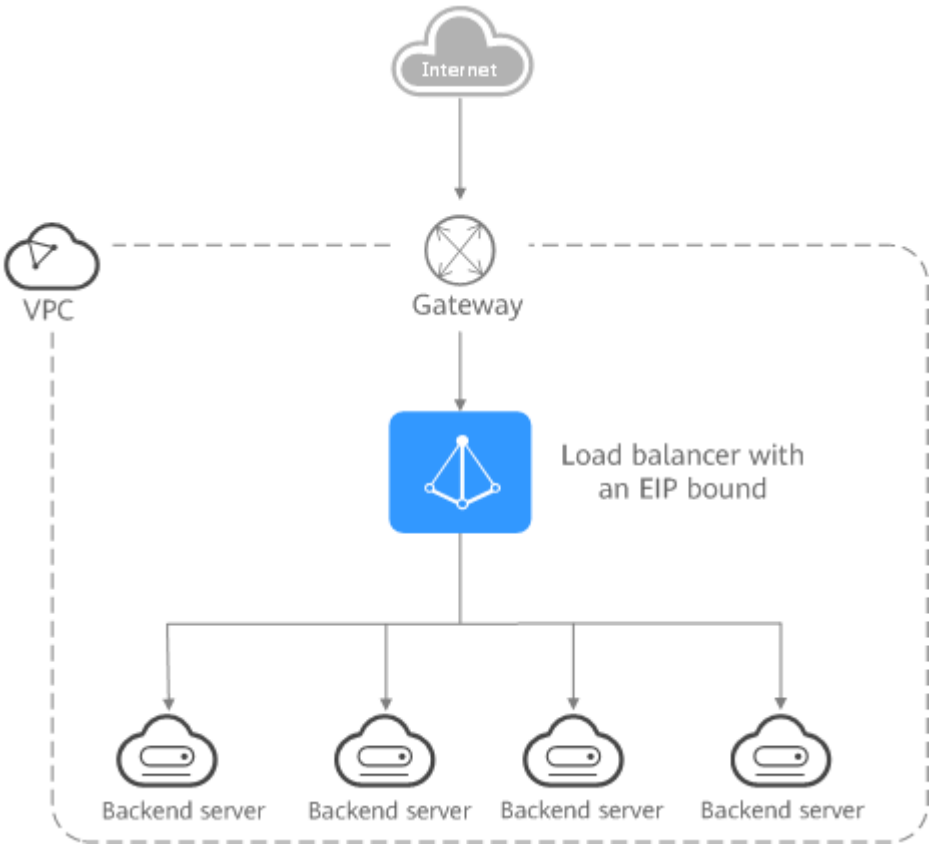
# 6 Equilibrio de carga en una red pública o privada

Un balanceador de carga puede funcionar en una red pública o privada.

## Equilibrio de carga en una red pública

Puede vincular un EIP a un balanceador de carga para que pueda recibir solicitudes de Internet y enrutar las solicitudes a los servidores backend.

**Figura 6-1** Equilibrio de carga en una red pública

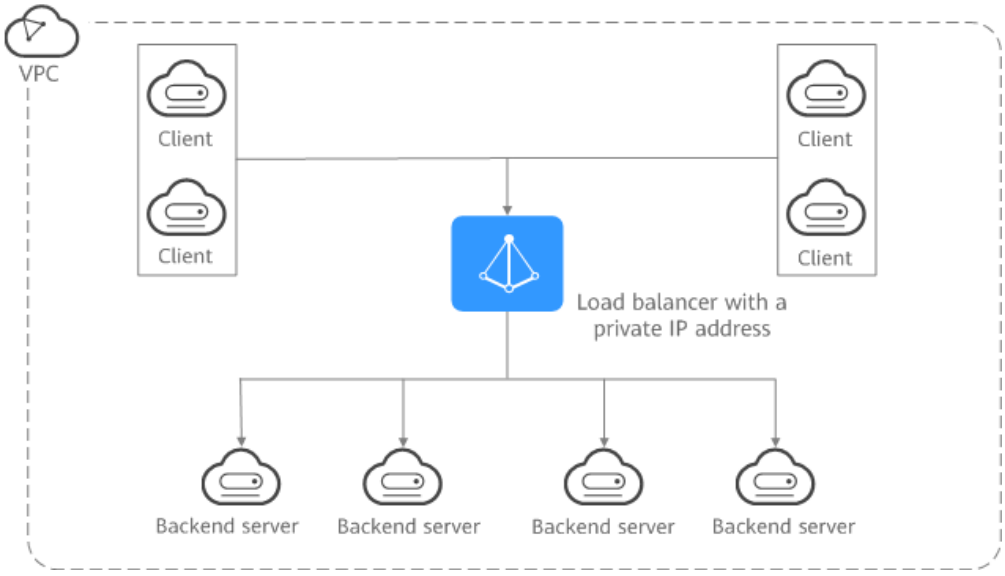


## Equilibrio de carga en una red privada

Un balanceador de carga solo tiene una dirección IP privada para recibir solicitudes de clientes en una VPC y enrutar las solicitudes a servidores backend en la misma VPC. Este tipo de balanceador de carga solo se puede acceder en una VPC.

**Figura 6-2** Equilibrio de carga en una red privada




## Tipos de red y balanceadores de carga

**Tabla 6-1** Balanceadores de carga dedicados y sus tipos de red

| Tipo de balanceador de carga | Tipo de red | Descripción |
|---|---|---|
| Balanceadores de carga dedicados | Red IPv4 pública | Cada balanceador de carga tiene un IPv4 EIP enlazado para permitirle enrutar solicitudes a través de Internet. |
| | Red IPv4 privada | Cada balanceador de carga solo tiene una dirección IPv4 privada y puede enrutar solicitudes en una VPC. |
| | Red IPv6 | Cada balanceador de carga tiene una dirección IPv6 enlazada.<br>● Si la dirección IPv6 se agrega a un ancho de banda compartido, el balanceador de carga puede enrutar solicitudes a través de Internet.<br>● Si la dirección IPv6 no se agrega a un ancho de banda compartido, el balanceador de carga puede enrutar solicitudes solo en una VPC. |

**Tabla 6-2** Balanceadores de carga compartidos y sus tipos de red

| Tipo de balanceador de carga | Tipo de red | Descripción |
| --- | --- | --- |
| Balanceadores de carga compartidos | Red IPv4 pública | Cada balanceador de carga tiene un EIP enlazado que le permite enrutar solicitudes a través de Internet. |
| | Red IPv4 privada | Cada balanceador de carga solo tiene una dirección IP privada y puede enrutar solicitudes en una VPC.<br>**NOTA**<br>Los balanceadores de carga compartidos admiten redes IPv4 privadas de forma predeterminada. La dirección IP privada de un balanceador de carga compartido no se puede cambiar. |

# 7 Rutas de tráfico de red

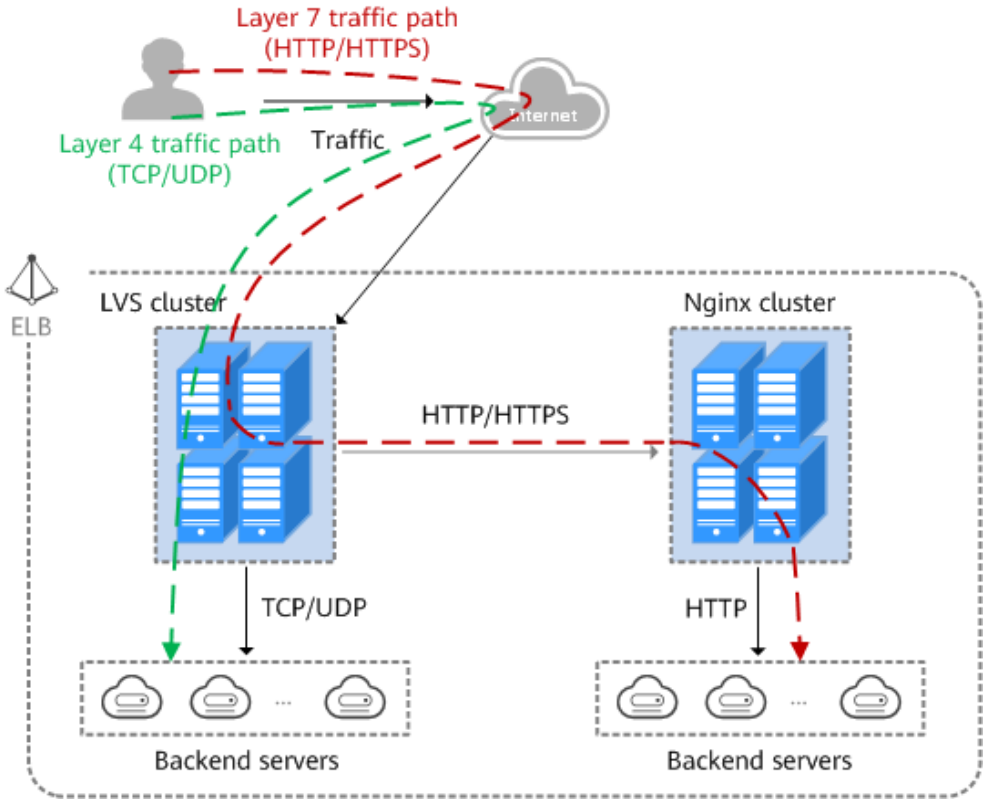
Los balanceadores de carga se comunican con servidores backend por una red privada.

- Si los servidores backend solo solicitan enrutado desde balanceadores de carga, no hay necesidad de asignar las EIP o crear los gateways de NAT.
- Si los servidores backend necesitan proporcionar servicios accesibles a Internet o acceder a Internet, debe asignar las EIP o crear los gateway de NAT.

## Rutas de tráfico de red entrante

Las configuraciones de los oyentes determinan cómo los balanceadores de carga distribuyen el tráfico entrante.

**Figura 7-1** Tráfico de red entrante



Cuando un oyente utiliza TCP o UDP para recibir tráfico entrante:

- El tráfico entrante solo se enruta a través del clúster LVS.
- El clúster de LVS enruta directamente el tráfico entrante a los servidores backend utilizando el algoritmo de equilibrio de carga que selecciona al agregar el oyente.

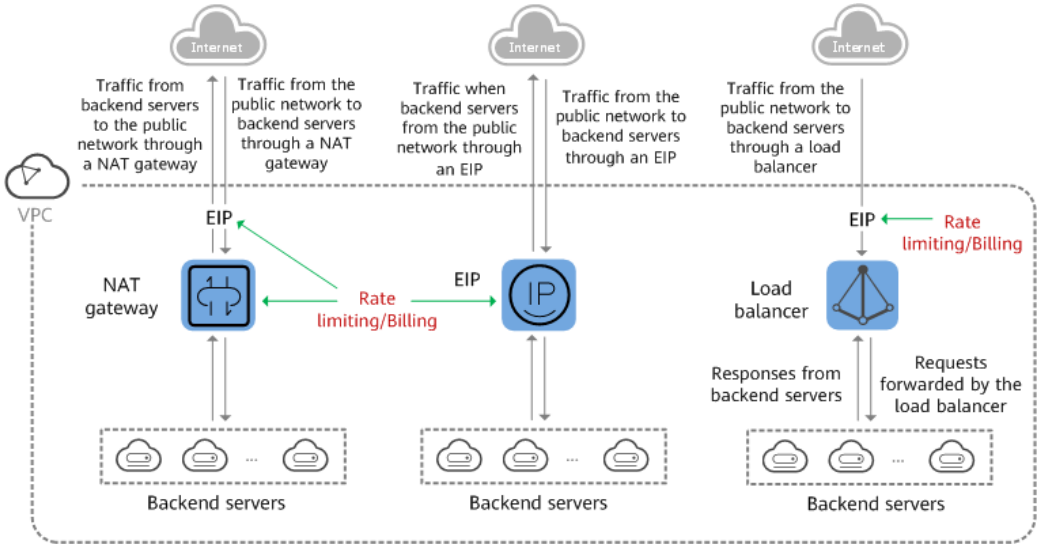
Cuando un oyente utiliza HTTP o HTTPS para recibir tráfico entrante:

- El tráfico entrante se enruta primero al clúster de LVS, luego al clúster de Nginx y, finalmente, a través de los servidores backend.
- Para el tráfico HTTPS, el clúster Nginx valida los certificados y descifra los paquetes de datos antes de distribuir el tráfico entre los servidores backend mediante HTTP.

## Rutas de tráfico de red saliente

El tráfico saliente se enruta de vuelta de la misma forma en que entró el tráfico.

**Figura 7-2** Tráfico de red saliente




- Debido a que el balanceador de carga recibe y responde a solicitudes a través de Internet, la transmisión de tráfico depende del ancho de banda, que no está limitado por ELB. El balanceador de carga se comunica con los servidores de backend a través de una red privada.
- Si tiene una puerta de gateway NAT, recibe y responde al tráfico entrante. El gateway NAT tiene un enlace EIP, a través del cual los servidores backend pueden acceder a Internet y proporcionar servicios accesibles desde Internet. Aunque hay una restricción en las conexiones que pueden ser procesadas por un gateway de NAT, la transmisión de tráfico depende del ancho de banda
- Si cada servidor backend tiene un EIP enlazado, reciben y responden al tráfico entrante directamente. La transmisión del tráfico depende del ancho de banda.

# 8 Especificaciones de los balanceadores de carga dedicados

Cuando crea un balanceador de carga dedicado, puede seleccionar especificaciones elásticas o fijas en función de sus requisitos de servicio. **Tabla 8-1** enumera las diferencias entre las dos especificaciones.

**Tabla 8-1** Comparación de especificaciones

| Concepto | Elástica | Fija |
| --- | --- | --- |
| Escenarios de aplicación | ● Para tráfico fluctuante<br>● Cuando necesite utilizar recursos temporalmente y con fines urgentes | ● Para un tráfico estable<br>● Cuando necesita usar recursos a largo plazo |
| Rendimiento del balanceador de carga de red (TCP/UDP/TLS) | El rendimiento se multiplica a medida que aumenta el número de AZ. **Tabla 8-2** muestra el rendimiento máximo en una AZ. | El rendimiento se multiplica a medida que aumenta el número de AZ. **Tabla 8-4** muestra el rendimiento máximo en una AZ. |
| Rendimiento del balanceador de carga de aplicaciones (HTTP/HTTPS) | El rendimiento se multiplica a medida que aumenta el número de AZ. **Tabla 8-2** muestra el rendimiento máximo en una AZ. | El rendimiento se multiplica a medida que aumenta el número de AZ. **Tabla 8-5** muestra el rendimiento máximo en una AZ. |
| Modo de facturación | Pago por uso | ● Pago por uso<br>● Anual/Mensual |
| Conceptos de facturación: | ● LCU<br>● Balanceador de carga | LCU |
| Capacidades | Iguales | |

## Especificaciones elásticas

Si su tráfico de servicio fluctúa considerablemente, puede elegir especificaciones elásticas y seleccionar el equilibrio de carga de red o aplicación que mejor se adapte a sus necesidades de servicio.

NOTA

El protocolo de oyente debe coincidir con el tipo de equilibrio de carga. Por ejemplo, si crea un balanceador de carga de aplicaciones, solo puede agregar un oyente de HTTP o de HTTPS a este balanceador de carga.

Los balanceadores de carga están disponibles en las diferentes especificaciones elásticas. Elija las especificaciones que mejor se adapten a sus necesidades. Cuando el tráfico excede lo definido en las especificaciones seleccionadas, se descartarán las nuevas solicitudes. Cada especificación tiene las siguientes dimensiones.

- Conexiones concurrentes máximas

  Indica el número máximo de conexiones simultáneas que un balanceador de carga puede manejar por minuto. Si el número alcanza el máximo de conexiones que se define en las especificaciones, se descartarán nuevas solicitudes para garantizar el rendimiento de las conexiones establecidas.

- Conexiones por segundo (CPS)

  Indica el número de conexiones nuevas que un balanceador de carga puede establecer por segundo. Si el número alcanza el CPS que se define en las especificaciones, se descartarán nuevas solicitudes para garantizar el rendimiento de las conexiones establecidas.

- Consultas por segundo (QPS)

  Indica el número de solicitudes HTTP o HTTPS enviadas a un servidor backend por segundo. Si el QPS alcanza lo definido en las especificaciones, se descartarán las nuevas solicitudes para garantizar el rendimiento de las conexiones establecidas.

- Ancho de banda (Mbit/s)

  Indica la cantidad máxima de datos que se pueden transmitir a través de una conexión por segundo.

**Tabla 8-2** Especificaciones elásticas máximas

| Protocolo | Conexiones simultáneas máximas | CPS | QPS | Ancho de banda (Mbit/s) |
|---|---|---|---|---|
| Balanceo de carga de red (TCP/UDP) | 20,000,000 | 400,000 | - | 10,000 |
| Equilibrio de carga de red (TLS) | 20,000,000 | 400,000 | - | 10,000 |
| Equilibrio de carga de aplicaciones (HTTP) | 8,000,000 | 80,000 | 160,000 | 10,000 |

| Protocolo | Conexiones simultáneas máximas | CPS | QPS | Ancho de banda (Mbit/s) |
|---|---|---|---|---|
| Equilibrio de carga de aplicaciones (HTTPS) | 8,000,000 | 80,000 | 160,000 | 10,000 |

**⚠ ATENCIÓN**

Las especificaciones elásticas disponibles se muestran en la consola y pueden variar según las regiones.

## Especificaciones fijas

Los balanceadores de carga están disponibles en las diferentes especificaciones fijas. Elija las especificaciones que mejor se adapten a sus necesidades. Cuando el tráfico excede lo definido en las especificaciones seleccionadas, se descartarán las nuevas solicitudes. Cada especificación tiene las siguientes dimensiones

- Conexiones concurrentes máximas

  Indica el número máximo de conexiones simultáneas que un balanceador de carga puede manejar por minuto. Si el número alcanza el máximo de conexiones que se define en **Tabla 8-4** y **Tabla 8-5**, se descartarán nuevas solicitudes para garantizar el rendimiento de las conexiones existentes.

- Conexiones por segundo (CPS)

  Indica el número de conexiones nuevas que un balanceador de carga puede establecer por segundo. Si el número llega al CPS que se define en **Tabla 8-4** y **Tabla 8-5**, se descartarán nuevas solicitudes para garantizar el rendimiento de las conexiones establecidas.

  Los oyentes de HTTPS necesitan crear acuerdos de enlace SSL para establecer conexiones con los clientes, y tales acuerdos de enlace SSL ocupan más recursos del sistema que los oyentes de HTTP. Por ejemplo, un balanceador de carga de aplicaciones small I puede establecer 2,000 nuevas conexiones de HTTP por segundo, pero solo 200 nuevas conexiones de HTTPS por segundo.

  Para un balanceador de carga de aplicaciones small I:

  – Si solo agrega un oyente de HTTP, el balanceador de carga puede establecer hasta 2,000 nuevas conexiones de HTTP.

  – Si solo agrega un oyente de HTTPS, el balanceador de carga puede establecer hasta 200 conexiones de HTTPS nuevas.

  – Si agregas un oyente de HTTPS y un oyente de HTTP, las nuevas conexiones se calculan utilizando la siguiente fórmula:

    Nuevas conexiones = Nuevas conexiones de HTTP + Nuevas conexiones de HTTPS x Relación de conexiones de HTTP a conexiones de HTTPS

Para un balanceador de carga de aplicaciones small I, la relación de conexiones de HTTP a conexiones de HTTPS es 10. Para obtener más información, consulte **Tabla 8-3**.

**Tabla 8-3** Nuevas conexiones que un balanceador de carga de aplicaciones small I puede establecer

| Parámetro | Escenario 1 | Escenario 2 |
| --- | --- | --- |
| Nuevas conexiones de HTTP | 1,000 | 1,000 |
| Nuevas conexiones de HTTPS | 50 | 150 |
| Nuevas conexiones de HTTP y de HTTPS | 1,000 + 50 x 10 = 1,500 | 1,000 + 150 x 10 = 2,500 |
| Descripción | ● Las nuevas conexiones no alcanzan el CPS (HTTP) definido en **Tabla 8-5** y las nuevas solicitudes se encaminarán correctamente. | ● Las nuevas conexiones llegan al CPS (HTTP) definido en **Tabla 8-5** y las nuevas solicitudes se encaminarán correctamente. |

NOTA

Los detalles en **Tabla 8-3** son solo para referencia.

- Consultas por segundo (QPS)

  Indica el número de solicitudes HTTP o HTTPS enviadas a un servidor backend por segundo. Si el QPS alcanza lo definido en **Tabla 8-5**, se descartarán nuevas solicitudes para garantizar el rendimiento de las conexiones establecidas.
- Ancho de banda (Mbit/s)

  Indica la cantidad máxima de datos que se pueden transmitir a través de una conexión por segundo.

**Tabla 8-4** y **Tabla 8-5** enumeran las especificaciones fijas de los balanceadores de carga dedicados.

ATENCIÓN

- **Las especificaciones fijas disponibles se muestran en la consola y pueden variar en función de los recursos de las distintas regiones.**
- El protocolo de oyente debe coincidir con el tipo de equilibrio de carga. Por ejemplo, si crea un balanceador de carga de aplicaciones, solo puede agregar un oyente de HTTP o de HTTPS a este balanceador de carga.

**Tabla 8-4** Especificaciones fijas para un balanceador de carga de red

| Tipo | Máximo de conexiones simultáneas | CPS | Ancho de banda (Mbit/s) | LCU en una AZ |
|---|---|---|---|---|
| Pequeño I | 500,000 | 10,000 | 50 | 10 |
| Pequeño II | 1,000,000 | 20,000 | 100 | 20 |
| Medio I | 2,000,000 | 40,000 | 200 | 40 |
| Medio II | 4,000,000 | 80,000 | 400 | 80 |
| Grande I | 10,000,000 | 200,000 | 1,000 | 200 |
| Grande II | 20,000,000 | 400,000 | 2,000 | 400 |

**Tabla 8-5** Especificaciones fijas para un balanceador de carga de aplicaciones

| Tipo | Máximo de conexiones simultáneas | CPS (HTTP) | CPS (HTTPS) | QPS (HTTP) | QPS (HTTPS) | Ancho de banda (Mbit/s) | LCU en una AZ |
|---|---|---|---|---|---|---|---|
| Pequeño I | 200,000 | 2,000 | 200 | 4,000 | 2,000 | 50 | 10 |
| Pequeño II | 400,000 | 4,000 | 400 | 8,000 | 4,000 | 100 | 20 |
| Medio I | 800,000 | 8,000 | 800 | 16,000 | 8,000 | 200 | 40 |
| Medio II | 2,000,000 | 20,000 | 2,000 | 40,000 | 20,000 | 400 | 100 |
| Grande I | 4,000,000 | 40,000 | 4,000 | 80,000 | 40,000 | 1,000 | 200 |
| Grande II | 8,000,000 | 80,000 | 8,000 | 160,000 | 80,000 | 2,000 | 400 |

NOTA

- Si agrega varios oyentes a un balanceador de carga, la suma de los valores QPS de todos los oyentes no puede exceder el QPS definido en cada especificación.
- El ancho de banda es el límite superior del tráfico entrante o saliente. Por ejemplo, para balanceadores de carga small I, el tráfico entrante o saliente no puede exceder los 50 Mbit/s.
- El ancho de banda incluido en cada especificación es el ancho de banda máximo proporcionado por ELB. Si se excede el ancho de banda máximo, el rendimiento de la red puede verse afectado.

# 9 Facturación (balanceadores de carga compartidos)

## Conceptos de facturación:

ELB admite el equilibrio de carga en redes públicas y privadas. **Tabla 9-1** describe los elementos de facturación.

**Tabla 9-1** Conceptos de facturación:

| Tipo de red | Balanceador de carga | EIP |
| --- | --- | --- |
| Red pública | Pago por uso | **Guía del usuario de Elastic IP** |
| Red privada | Pago por uso | No incluidos |

NOTA

- Si sus balanceadores de carga compartidos se crearon después del 10 de febrero de 2023, el rendimiento garantizado estaba habilitado para ellos de forma predeterminada, y los balanceadores de carga se facturarán de forma de pago por uso. Los balanceadores de carga compartidos creados antes del 10 de febrero de 2023 son gratuitos porque el rendimiento garantizado no está habilitado para ellos. Puede habilitar el rendimiento garantizado haciendo referencia a *Garantizar el rendimiento de un balanceador de carga compartido*. Una vez que esta función esté activada, sus balanceadores de carga se cargarán sobre la base de pago por uso.
- El pago por uso es el único modo de facturación para los balanceadores de carga. Si vincula un nuevo EIP a un balanceador de carga cuando crea el balanceador de carga, el modo de facturación de este EIP es de pago por uso de forma predeterminada. Si necesita un EIP anual/mensual, puede comprar un EIP anual/mensual por separado y vincularlo al balanceador de carga cuando cree el balanceador de carga.
- Para obtener más información, consulte **Detalles de precios de ELB**.

## Modo de facturación

Los balanceadores de carga compartidos con rendimiento garantizado se facturan en función de pago por uso. Cada balanceador de carga en una red pública tiene una EIP vinculada para recibir solicitudes de Internet.

El ancho de banda que utilizará el EIP se factura por tráfico o ancho de banda fijo:

- Por tráfico: especifica un ancho de banda máximo y paga por el tráfico saliente total. Esto es adecuado para aplicaciones con picos y valles predecibles en el tráfico.
- Por ancho de banda: especifica un ancho de banda máximo y paga por la cantidad de tiempo que usa el ancho de banda. Esto es ideal para aplicaciones que requieren un ancho de banda estable.

## Cambio de la configuración de ancho de banda

Puede cambiar la configuración de ancho de banda, incluidos su nombre, tamaño y opción de facturación.

Si cambia la opción de facturación a **By bandwidth**, se le cobrará el tiempo que uses el ancho de banda. Si cambia la opción de facturación a **By traffic**, se le cobrará el total del tráfico saliente.

Si su cuenta permanece en mora después de que finalice el período de gracia, los balanceadores de carga no se pueden usar para recibir solicitudes de Internet. Todavía puede agregar oyentes y asociar servidores backend con los balanceadores de carga.

# 10 Permisos

Si necesita asignar diferentes permisos al personal de su empresa para acceder a sus recursos de ELB, IAM es una buena opción para la gestión de permisos detallada. IAM proporciona autenticación de identidad, gestión de permisos y control de acceso, lo que le ayuda a acceder de forma segura a sus recursos en la nube.

Con IAM, puede crear usuarios de IAM y asignar permisos para controlar su acceso a recursos específicos. Por ejemplo, si desea que algunos desarrolladores de software de su empresa utilicen recursos de ELB pero no desea que eliminen estos recursos ni realicen otras operaciones de alto riesgo, puede conceder permiso para usar recursos de ELB pero no permiso para eliminarlos.

Omita esta sección si su cuenta de Huawei Cloud no requiere usuarios individuales de IAM para la gestión de permisos.

IAM es un servicio gratuito. Solo paga por los recursos de su cuenta. Para obtener más información acerca de IAM, consulte **Descripción del servicio de IAM**.

## Permisos de ELB

Los nuevos usuarios de IAM no tienen ningún permiso asignado de forma predeterminada. Primero debe agregarlos a uno o más grupos y adjuntar políticas o roles a estos grupos. A continuación, los usuarios heredan los permisos de los grupos y pueden realizar operaciones específicas en servicios en la nube en función de los permisos que se les han asignado.

ELB es un servicio a nivel de proyecto desplegado para regiones específicas. Para asignar permisos ELB a un grupo de usuarios, especifique el ámbito como proyectos específicos de la región y seleccione los proyectos para los que desea que los permisos surtan efecto. Si selecciona **All projects**, los permisos surtirán efecto para el grupo de usuarios en todos los proyectos específicos de la región. Al acceder a ELB, los usuarios deben cambiar a la región autorizada.

Puede conceder permisos a los usuarios mediante roles y políticas.

- Roles: Una estrategia de autorización de grano grueso proporcionada por IAM para asignar permisos en función de las responsabilidades del trabajo de los usuarios. Solo un número limitado de roles de nivel de servicio están disponibles para autorización. Cuando concede permisos mediante roles, también debe adjuntar las dependencias de roles existentes. Los roles no son ideales para la autorización detallada y el acceso con privilegios mínimos.

- Políticas: Una estrategia de autorización detallada proporcionada por IAM para asignar los permisos necesarios para realizar operaciones en recursos específicos en la nube bajo ciertas condiciones. Este tipo de autorización es más flexible y es ideal para el acceso de privilegios mínimos. Por ejemplo, puede conceder a los usuarios de ELB solo permisos para gestionar un determinado tipo de recursos. La mayoría de las políticas detalladas contienen permisos para API específicas, y los permisos se definen mediante acciones de API. Para ver las acciones de API admitidas por ELB, consulta **Políticas de permisos y acciones admitidas**.

**Tabla 10-1** enumera todos los permisos definidos por el sistema para ELB.

**Tabla 10-1** Permisos definidos por el sistema para ELB

| Nombre de rol/política | Descripción | Tipo |
|---|---|---|
| ELB FullAccess | Permisos: todos los permisos en los recursos ELB<br>Alcance: servicio a nivel de proyecto | Política definida por el sistema |
| ELB ReadOnlyAccess | Permisos: permisos de solo lectura en recursos ELB<br>Alcance: servicio a nivel de proyecto | Política definida por el sistema |
| ELB Administrator | Permisos: todos los permisos en los recursos ELB. Para obtener este permiso, los usuarios también deben tener los permisos **Tenant Administrator**, **VPC Administrator**, **CES Administrator**, **Server Administrator** y **Tenant Guest**.<br>Alcance: servicio a nivel de proyecto<br>**NOTA**<br>Si su cuenta ha solicitado permisos detallados, configure políticas detalladas para los permisos del sistema de ELB, en lugar de las políticas de administrador de ELB. | Rol definido por el sistema |

**Tabla 10-2** describe las operaciones comunes soportadas por cada política de sistema de ELB.

**Tabla 10-2** Operaciones comunes respaldadas por políticas definidas por el sistema

| Operación | ELB FullAccess | ELB ReadOnlyAccess | Administrador de ELB |
|---|---|---|---|
| Crear un balanceador de carga | Se admite | No se admite | Se admite |
| Consultar un balanceador de carga | Se admite | Se admite | Se admite |

| Operación | ELB FullAccess | ELB ReadOnlyAccess | Administrador de ELB |
|---|---|---|---|
| Consultar un balanceador de carga y recursos asociados | Se admite | Se admite | Se admite |
| Consultar los balanceadores de carga | Se admite | Se admite | Se admite |
| Modificación de un balanceador de carga | Se admite | No se admite | Se admite |
| Eliminar un balanceador de carga | Se admite | No se admite | Se admite |
| Agregar un oyente | Se admite | No se admite | Se admite |
| Consultar un oyente | Se admite | Se admite | Se admite |
| Modificar un oyente | Se admite | No se admite | Se admite |
| Eliminar un oyente | Se admite | No se admite | Se admite |
| Agregar un grupo de servidores de backend | Se admite | No se admite | Se admite |
| Consultar un grupo de servidores backend | Se admite | Se admite | Se admite |
| Modificar un grupo de servidores backend | Se admite | No se admite | Se admite |
| Eliminar un grupo de servidores backend | Se admite | No se admite | Se admite |
| Agregar un servidor de backend | Se admite | No se admite | Se admite |
| Consultar un servidor de backend | Se admite | Se admite | Se admite |
| Modificar un servidor de backend | Se admite | No se admite | Se admite |
| Eliminar un servidor de backend | Se admite | No se admite | Se admite |
| Configurar una comprobación de estado | Se admite | No se admite | Se admite |
| Consultar una comprobación de estado | Se admite | Se admite | Se admite |
| Modificar una comprobación de estado | Se admite | No se admite | Se admite |

| Operación | ELB FullAccess | ELB ReadOnlyAccess | Administrador de ELB |
|---|---|---|---|
| Deshabilitar una comprobación de estado | Se admite | No se admite | Se admite |
| Asignar un EIP | No se admite | No se admite | Se admite |
| Vincular un EIP a un balanceador de carga | No se admite | No se admite | Se admite |
| Consultar un EIP | Se admite | Se admite | Se admite |
| Desvincular un EIP de un balanceador de carga | No se admite | No se admite | Se admite |
| Consultar Métricas | No se admite | No se admite | Se admite |
| Consultar logs de acceso | No se admite | No se admite | Se admite |

NOTA

- Para desvincular un EIP, también necesita configurar los permisos **vpc:bandwidths:update** y **vpc:publicIps:update** del servicio VPC. Para obtener más información, consulta la *Referencia de la API de Virtual Private Cloud*.
- Para ver las métricas de supervisión, también debe configurar el permiso **CES ReadOnlyAccess**. Para obtener más información, consulta la *Referencia de la API de Cloud Eye*.
- Para ver los registros de acceso, también necesita configurar el permiso **LTS ReadOnlyAccess**. Para obtener más información, consulta la *Referencia de la API del Log Tank Service*.

# 11 Conceptos de producto

## 11.1 Conceptos básicos

**Tabla 11-1** Algunos conceptos sobre ELB

| Término | Definición |
| --- | --- |
| Balanceador de carga | Un balanceador de carga distribuye el tráfico entrante entre los servidores backend. |
| Oyente | Un oyente escucha las peticiones de los clientes y dirige las peticiones al backend basándose en las opciones que se configuran al agregar el receptor. |
| Servidor backend | Un servidor backend es un servidor en la nube agregado a un grupo de backend asociado con un balanceador de carga. Cuando agrega un oyente a un balanceador de carga, puede crear o seleccionar un grupo de backend para recibir solicitudes del balanceador de carga mediante el puerto y el protocolo que especifique para el grupo de backend y el algoritmo de balanceo de carga que seleccione. |
| Grupo de servidores backend | Un grupo de servidores backend es una colección de servidores en la nube que tienen las mismas características. Cuando se agrega un oyente, se selecciona un algoritmo de equilibrio de carga y se crea o se selecciona un grupo de servidores backend. El tráfico entrante se enruta al grupo de servidores backend correspondiente en función de la configuración del oyente. |
| Comprobación de estado | ELB envía periódicamente solicitudes a los servidores de backend para comprobar si pueden procesar solicitudes. Si se detecta un backend como no saludable, el balanceador de carga detiene las solicitudes de enrutamiento a él. Después de que el backend se recupere, el balanceador de carga reanudará las solicitudes de enrutamiento a él. |
| Redireccionamiento | HTTPS es una extensión de HTTP. HTTPS cifra los datos entre un servidor web y un navegador. |

| Término | Definición |
| --- | --- |
| Sesión persistente | Las sesiones persistentes garantizan que las solicitudes de un cliente siempre se enruten al mismo servidor back-end antes de que transcurra una sesión. |
| WebSocket | WebSocket es un nuevo protocolo HTML5 que proporciona comunicación full-duplex entre el navegador y el servidor. WebSocket ahorra recursos del servidor y ancho de banda, y permite la comunicación en tiempo real. Tanto WebSocket como HTTP dependen de TCP para transmitir datos. Se requiere una conexión de protocolo de enlace entre el navegador y el servidor, para que puedan comunicarse entre sí solo después de que se establezca la conexión. Sin embargo, como protocolo de comunicación bidireccional, el WebSocket es diferente de HTTP. Después de que el protocolo de enlace tenga éxito, tanto el servidor como el navegador (o agente cliente) pueden enviar datos o recibir datos entre sí de forma activa. |
| SNI | SNI, una extensión de Transport Layer Security (TLS), permite a un servidor presentar varios certificados en la misma dirección IP y número de puerto. SNI permite al cliente indicar el nombre de dominio del sitio web mientras envía una solicitud de protocolo de enlace SSL. Una vez que recibe la solicitud, el balanceador de carga consulta el certificado correcto basado en el nombre de host o el nombre de dominio y devuelve el certificado al cliente. Si no se encuentra ningún certificado, el balanceador de carga devolverá el certificado predeterminado. |
| Conexión persistente | Una conexión persistente permite que múltiples paquetes de datos se envíen continuamente a través de una conexión TCP. Si no se envía ningún paquete de datos durante la conexión, el cliente y el servidor envían paquetes de detección de enlace entre sí para mantener la conexión. |
| Conexión corta | Una conexión corta es una conexión establecida cuando se intercambian datos entre el cliente y el servidor e inmediatamente se cierra después de que se envían los datos. |
| Conexión simultánea | Las conexiones simultáneas son el número total de conexiones TCP iniciadas por los clientes y enrutadas a los servidores backend por un balanceador de carga por segundo. |

# 11.2 Región y AZ

## Concepto

Una región y una zona de disponibilidad (AZ) identifican la ubicación de un centro de datos. Puede crear recursos en una región específica y AZ.

- Las regiones se dividen en función de la ubicación geográfica y la latencia de la red. Los servicios públicos, como Elastic Cloud Server (ECS), Elastic Volume Service (EVS), Object Storage Service (OBS), Virtual Private Cloud (VPC), Elastic IP (EIP) y Image Management Service (IMS), se comparten dentro de la misma región. Las regiones se clasifican en regiones universales y regiones dedicadas. Una región universal

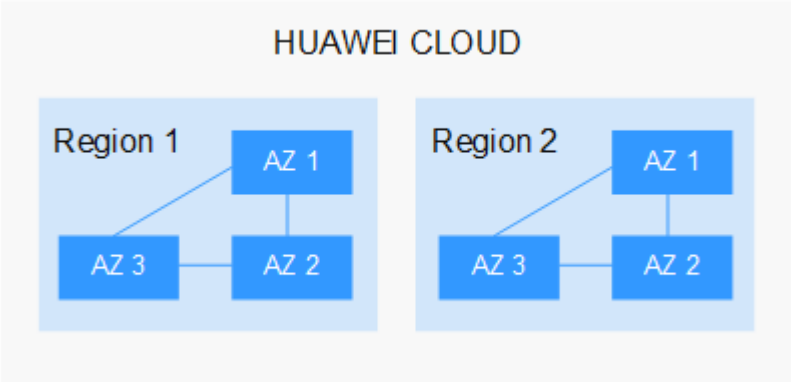
  

proporciona servicios en la nube universales para los tenants estándares. Una región dedicada proporciona servicios específicos para tenants específicos.

- Una AZ contiene uno o más centros de datos físicos. Cada AZ cuenta con instalaciones independientes de electricidad, de refrigeración, de extinción de incendios y a prueba de humedad. Dentro de una AZ, los recursos de computación, red, almacenamiento y otros se dividen de forma lógica en múltiples clústeres. Las AZ dentro de una región están interconectadas usando fibras ópticas de alta velocidad, para soportar sistemas de alta disponibilidad entre las AZ.

**Figura 11-1** muestra la relación entre regiones y AZ.

**Figura 11-1** Las regiones y las AZ




Huawei Cloud ofrece servicios en muchas regiones de todo el mundo. Seleccione una región y AZ según los requisitos. Para obtener más información, consulte **Regiones globales de Huawei Cloud**.

## Selección de una región

Al seleccionar una región, tenga en cuenta los siguientes factores:

- Localización

  Se recomienda seleccionar la región más cercana para una menor latencia de red y un acceso rápido. Las regiones dentro de China continental proporcionan la misma infraestructura, calidad de red BGP, así como operaciones de recursos y configuraciones. Por lo tanto, si sus usuarios objetivo están en China continental, no es necesario tener en cuenta las diferencias de latencia de la red al seleccionar una región.

  – Si sus usuarios objetivo se encuentran en Asia Pacífico (excepto China continental), seleccione la región **CN-Hong Kong**, **AP-Bangkok**, or **AP-Singapore**.
  – Si sus usuarios objetivo se encuentran en África, seleccione la región **AF-Johannesburg**.
  – Si sus usuarios objetivo están en América Latina, seleccione la región **LA-Santiago**.

  NOTA

  La región **LA-Santiago** se encuentra en Chile.

- Precio del recurso

  Los precios de los recursos pueden variar en diferentes regiones. Para obtener más información, consulte **Detalles de precios del producto**.

## Selección de una AZ

Al implementar recursos, tenga en cuenta los requisitos de las aplicaciones en cuanto a la recuperación ante desastres (DR) y la latencia de la red.

- Para una alta capacidad de DR, implemente recursos en diferentes AZ dentro de la misma región.
- Para una menor latencia de red, implemente recursos en la misma AZ.

## Regiones y endpoint

Antes de usar una API para llamar a recursos, especifique su región y endpoint.

# 12 Cómo funciona ELB con otros servicios

**Tabla 12-1** Servicios relacionados

| Nombre del servicio | Función | Referencia |
|---|---|---|
| Elastic Cloud Server (ECS) | Proporciona servidores para ejecutar sus aplicaciones en la nube. Configure los balanceadores de carga para enrutar el tráfico a los servidores o contenedores. | **Compra e inicio de sesión en un ECS de Linux** |
| Bare Metal Server (BMS) | | **Creación de un BMS** |
| Virtual Private Cloud (VPC) | Proporciona direcciones IP y ancho de banda para balanceadores de carga. | **Asignación de una EIP** |
| Auto Scaling (AS) | Funciona con ELB para escalar automáticamente el número de servidores backend para una distribución más rápida del tráfico. | **Creación de un grupo de AS** |
| Identity and Access Management (IAM) | Proporciona la autenticación para ELB. | **Creación de un grupo de usuarios y asignación de permisos** |
| Cloud Trace Service (CTS) | Registra las operaciones realizadas en recursos de ELB. | **Consulta de trazas** |
| Cloud Eye | Supervisa el estado de los balanceadores de carga y oyentes, sin ningún complemento adicional. | **Consulta de Métricas** |
| Anti-DDoS | Defiende los balanceadores de carga de redes públicas contra ataques de DDoS, manteniendo su negocio estable y confiable. | **Configuración de una política de protección anti-DDoS** |